

Analysis of Longitudinal Data in Perinatal Trials when the Length of Follow-up is Informative

Tessa Longstaff

Thesis submitted for the degree of

Master of Philosophy

in

Statistics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences



THE UNIVERSITY
of ADELAIDE

May 14, 2016

Contents

Signed Statement	xiv
Signed Statement	xv
Acknowledgements	xvi
Abstract	xvii
1 Introduction	1
1.1 Motivating Example	1
1.2 Research Aims	2
1.3 Thesis Outline	3
2 Methods for Analysing Clustered Data	4
2.1 Clustered Data	4
2.2 Standard Methods of Analysis for Clustered Data	5
2.2.1 Mixed Model	5
2.2.2 Generalised Estimating Equations	7
2.2.3 Comparison	11
2.3 Informative Cluster Size	12
2.3.1 Definition	12
2.3.2 Examples	12
2.3.3 Issues Arising for Data with ICS	13
2.3.4 Relationship with Missing Data	15
2.4 Approaches for Dealing with ICS	15

2.4.1	Within Cluster Resampling	16
2.4.2	Joint Models	17
2.4.3	Include Cluster Size as a Covariate	18
2.4.4	GEE Based Approaches	18
2.4.5	Approaches for the POPPET Trial	22
2.5	Summary of Methods	22
3	POPJET Data Analysis	23
3.1	Variables in the POPJET Data Set	25
3.2	ICS for the POPJET Data	25
3.3	Mixed Model Analysis	27
3.3.1	Model Development	27
3.3.2	Informative Cluster Size	32
3.3.3	Reducing Informative Cluster Size by Adjusting for Baseline Co- variates	33
3.3.4	Mixed Model Analysis Results	36
3.3.5	Attempt to Remove ICS by Having a Fixed Trial Length	41
3.4	GEE Analysis	43
3.5	Comparison of Methods	51
4	Simulation Study	54
4.1	Generating Simulated Data	54
4.1.1	Gestational Age in Weeks	55
4.1.2	Birthweight	55
4.1.3	Treatment Group	57
4.1.4	Coefficients, Error and Random Effects	58
4.1.5	Infant Weights Over Time	59
4.2	Method for Determining Cluster Size	59
4.2.1	Expected Value and Variance of Cluster Size Distribution	60
4.2.2	Determine Parameters to Match POPJET Trial Cluster Size Dis- tribution	63
4.3	Simulation Scenarios	65

4.4	Results	68
4.4.1	NICS Scenarios	68
4.4.2	ICS Scenarios	72
4.5	Simulation Study Conclusions	78
5	Simulation Study Extensions	79
5.1	Extension 1: Larger Standard Deviation of Cluster Size	80
5.1.1	Results for Extension 1	82
5.2	Extension 2: Larger Sample Size	87
5.2.1	NICS Scenario Results for Extension 2	88
5.2.2	ICS Scenario Results for Extension 2	91
5.3	Extension 3: Simulating from a GEE Framework	102
5.3.1	Generating Simulated Data from a GEE Framework	103
5.3.2	NICS Scenario Results for Extension 3	108
5.3.3	ICS Scenario Results for Extension 3	113
5.4	Extensions Conclusions	124
5.5	Implications for the POPPET data	125
6	Conclusion	129
6.1	Main Findings	129
6.2	Significance and Innovation	132
6.3	Limitations and Future Work	133
6.4	Final Recommendations	133
A	Results	134
A.1	Chapter 4 Simulation Results	134
A.1.1	Scenarios with Non Informative Cluster Size	134
A.1.2	Scenarios with Informative Cluster Size using $\gamma_0 = 3.069142$	137
A.2	Extension 1: Simulation With Larger Standard Deviation of Cluster Size, $\gamma_0 = 2.787813$	142
B	Results Plots	148
B.1	Plots for Chapter 4 Simulation Study Results	148

B.2	Extension 1: Plots for Larger Standard Deviation of Cluster Size	155
B.3	Extension 2: Plots for Larger Sample Size Results	160
B.4	Extension 3: Plots for GEE Study Results	173
C	Results for Larger Trial (600 individuals)	188
C.1	Scenarios with Non Informative Cluster Size	188
C.2	Scenarios with smaller SD of Cluster Size, $\gamma_0 = 3.069142$	191
C.3	Scenarios with Larger SD of Cluster Size, $\gamma_0 = 2.787813$	196
D	GEE Results	202
D.1	GEE Results for n=60	202
D.1.1	Scenarios with Non Informative Cluster Size	202
D.1.2	Scenarios with Informative Cluster Size using $\gamma_0 = 3.069142$	205
D.2	GEE Results for n=600	210
D.2.1	Scenarios with Non Informative Cluster Size	210
D.2.2	Scenarios with Informative Cluster Size using $\gamma_0 = 3.069142$	213

List of Tables

2.1	Coefficients for the IEE and EGEE methods	14
2.2	Probability of disease for the IEE and EGEE methods	14
3.1	Variables to be used in the analysis of the POPPET data set	25
3.2	Summary of initial duration model	34
3.3	Summary of final duration model	35
3.4	Summary of fixed effects for mixed model	37
3.5	Summary of random effects for mixed model	37
3.6	Summary of fixed effects for fixed trial length model	42
3.7	Summary of random effects for fixed trial length model	42
3.8	Summary of the EGEE model	44
3.9	Summary of the ARGEE model	44
3.10	Summary of the IEE model	45
3.11	Summary of the CWGEE model	45
3.12	<i>Group</i> \times <i>Time</i> estimates for each model	52
3.13	<i>Group</i> \times <i>Time</i> ² estimates for each model	52
3.14	Trajectories for an infant given each group allocation	53
4.1	Summary of fixed effects for simulation study mixed model	58
4.2	Summary of random effects for simulation study mixed model	58
4.3	Values of γ_1 and γ_2 and resulting correlations	65
4.4	Simulation Scenarios	68
4.5	Scenario 1 (fixed trial length of 38 days, treatment effect = 28 g/week)	69
4.6	Scenario 2 ($\gamma_0 = 3.19, \gamma_1 = 0, \gamma_2 = 0$, treatment effect = 28 g/week)	69
4.7	Coverage probabilities for NICS scenarios	71

4.8	Scenario 3 ($\gamma_1 = -0.50, \gamma_2 = 0$, treatment effect = 28 g/week)	72
4.9	Scenario 4 ($\gamma_1 = 0, \gamma_2 = -0.50$, treatment effect = 28 g/week)	72
4.10	Scenario 5 ($\gamma_1 = -0.35, \gamma_2 = -0.35$, treatment effect = 28 g/week)	72
4.11	Scenario 6 ($\gamma_1 = -0.46, \gamma_2 = -0.19$, treatment effect = 28 g/week)	73
4.12	Scenario 7 ($\gamma_1 = -0.19, \gamma_2 = -0.46$, treatment effect = 28 g/week)	73
4.13	Ratio of mean SE divided by SD	75
4.14	Coverage probabilities for ICS scenarios	77
5.1	Extension 1: Values of γ_1 and γ_2 and resulting correlations	80
5.2	Simulation Scenarios for Extension 1	81
5.3	Scenario 36 ($\gamma_1=-0.90, \gamma_2=0$, treatment effect = 28 g/week)	82
5.4	Scenario 37 ($\gamma_2=-0.90, \gamma_1=0$, treatment effect = 28 g/week)	82
5.5	Scenario 38 ($\gamma_1=-0.64, \gamma_2=-0.64$, treatment effect = 28 g/week)	82
5.6	Scenario 39 ($\gamma_1=-0.83, \gamma_2=-0.34$, treatment effect = 28 g/week)	82
5.7	Scenario 40 ($\gamma_2=-0.83, \gamma_1=-0.34$, treatment effect = 28 g/week)	83
5.8	Extension 1: Ratios of mean SE divided by SD	85
5.9	Extension 1: Coverage probabilities	87
5.10	Extension 2 Scenario 1 (fixed trial length=38, treatment effect = 28 g/week)	88
5.11	Extension 2 Scenario 2 ($\gamma_0=\log(24.3667), \gamma_1=0, \gamma_2=0$, treatment effect = 28 g/week)	88
5.12	Extension 2: Coverage probabilities for NICS scenarios	91
5.13	Extension 2 Scenario 3 ($\gamma_1=-0.50, \gamma_2=0$, treatment effect = 28 g/week) . .	91
5.14	Extension 2 Scenario 4 ($\gamma_2=-0.50, \gamma_1=0$, treatment effect = 28 g/week) . .	92
5.15	Extension 2 Scenario 5 ($\gamma_1=-0.35, \gamma_2=-0.35$, treatment effect = 28 g/week)	92
5.16	Extension 2 Scenario 6 ($\gamma_1=-0.46, \gamma_2=-0.19$, treatment effect = 28 g/week)	92
5.17	Extension 2 Scenario 7 ($\gamma_2=-0.46, \gamma_1=-0.19$, treatment effect = 28 g/week)	92
5.18	Extension 2 Scenario 36 ($\gamma_1=-0.90, \gamma_2=0$, treatment effect = 28 g/week) . .	93
5.19	Extension 2 Scenario 37 ($\gamma_2=-0.90, \gamma_1=0$, treatment effect = 28 g/week) . .	93
5.20	Extension 2 Scenario 38 ($\gamma_1=-0.64, \gamma_2=-0.64$, treatment effect = 28 g/week)	93
5.21	Extension 2 Scenario 39 ($\gamma_1=-0.83, \gamma_2=-0.34$, treatment effect = 28 g/week)	93
5.22	Extension 2 Scenario 40 ($\gamma_2=-0.83, \gamma_1=-0.34$, treatment effect = 28 g/week)	94

5.23	Extension 2: Ratios of mean SE divided by SD using smaller SD of cluster size	99
5.24	Extension 2: Ratios of mean SE divided by SD using larger SD of cluster size	100
5.25	Extension 2: Coverage probabilities for smaller SD of cluster size	101
5.26	Extension 2: Coverage probabilities for larger SD of cluster size	102
5.27	Extension 3: Values of γ_1 and γ_2 and resulting correlations	108
5.28	Extension 3 Scenario 1 (fixed trial length=38, treatment effect = 28 g/week)	108
5.29	Extension 3 Scenario 2 ($\gamma_0=\log(24.3667), \gamma_1=0, \gamma_2=0$, treatment effect = 28 g/week)	109
5.30	Extension 3 (Larger Trial) Scenario 1 (fixed trial length=38, treatment effect = 28 g/week)	109
5.31	Extension 3 (Larger Trial) Scenario 2 ($\gamma_0=\log(24.3667), \gamma_1=0, \gamma_2=0$, treatment effect = 28 g/week)	109
5.32	Extension 3: Coverage probabilities for NICS scenarios for n=60	113
5.33	Extension 3: Coverage probabilities for NICS scenarios for n=600	113
5.34	Extension 3 Scenario 3 ($\gamma_1=-0.36, \gamma_2=-0.22$, treatment effect = 28 g/week)	114
5.35	Extension 3 Scenario 4 ($\gamma_1=-0.039, \gamma_2=-0.48$, treatment effect = 28 g/week)	114
5.36	Extension 3 Scenario 5 ($\gamma_1=-0.24, \gamma_2=-0.35$, treatment effect = 28 g/week)	114
5.37	Extension 3 Scenario 6 ($\gamma_1=-0.29, \gamma_2=-0.29$, treatment effect = 28 g/week)	114
5.38	Extension 3 Scenario 7 ($\gamma_1=-0.17, \gamma_2=-0.40$, treatment effect = 28 g/week)	115
5.39	Extension 3 (Larger Trial) Scenario 3 ($\gamma_1=-0.36, \gamma_2=-0.22$, treatment effect = 28 g/week)	115
5.40	Extension 3 (Larger Trial) Scenario 4 ($\gamma_1=-0.039, \gamma_2=-0.48$, treatment effect = 28 g/week)	115
5.41	Extension 3 (Larger Trial) Scenario 5 ($\gamma_1=-0.24, \gamma_2=-0.35$, treatment effect = 28 g/week)	115
5.42	Extension 3 (Larger Trial) Scenario 6 ($\gamma_1=-0.29, \gamma_2=-0.29$, treatment effect = 28 g/week)	116
5.43	Extension 3 (Larger Trial) Scenario 7 ($\gamma_1=-0.17, \gamma_2=-0.40$, treatment effect = 28 g/week)	116

5.44	Extension 3: Bias amount for n=60	118
5.45	Extension 3: Bias amount for n=600	119
5.46	Extension 3: Ratios of mean SE divided by SD for n=60	120
5.47	Extension 3: Ratios of mean SE divided by SD for n=600	120
5.48	Extension 3: Coverage probabilities for n=60	123
5.49	Extension 3: Coverage probabilities for n=600	123
5.50	Results for the POPPET data for a MM	127
5.51	Results for the POPPET data for a ARGEE	128

List of Figures

3.1	Scatterplot of weight against time for each infant	27
3.2	Plots of main group effect vs group x time interaction, with no main effect	29
3.3	Scatterplot of residuals vs fitted values	30
3.4	Scatterplot of residuals vs fitted values for quadratic model	30
3.5	Normal quantile plot of residuals	31
3.6	Normal quantile plot of random effects	31
3.7	Scatterplot of random effects vs duration	33
3.8	Scatterplot of random effects vs duration for model with baseline covariates	36
3.9	Actual (in black) and expected (in red) weight against time for each infant	38
3.10	Scatterplot of weight against time for every 10th infant with a plot of fitted values against time for every 10th infant (in red)	39
3.11	Scatterplot of mean weight for Group 1 (black), mean weight for Group 2 (blue) and mean fixed effects for each group (Group 1 is green and Group 2 is red) vs time	40
3.12	Scatterplot of mean weight for Group 1 (black), mean weight for Group 2 (blue) and mean fixed effects for each group (Group 1 is green and Group 2 is red) vs time with a plot of the fitted values for each infant for the entire duration of 72 days. It is pink while the infant was in the study and then light green once the infant was discharged.	41
3.13	Actual (in black) and expected (in red) weight against time for each infant for the EGEE	47
3.14	Actual (in black) and expected (in red) weight against time for each infant for the ARGEE	48

3.15	Actual (in black) and expected (in red) weight against time for each infant for the IEE	48
3.16	Actual (in black) and expected (in red) weight against time for each infant for the CWGEE	49
3.17	Actual (in black) and expected (in red) weight against time for each infant for the MM	49
3.18	Scatterplot of mean weight for Group 1 (black), for Group 2 (light green) and mean fitted values from GEE for each group (Group 1 is blue and Group 2 is pink) vs time for each of the GEEs	51
4.1	Histogram of gestational age in weeks	56
4.2	Histogram of birthweight in grams	56
4.3	Scatterplot of birthweight vs gestational age	57
4.4	Histogram of duration	60
4.5	Boxplots of Interaction Effect estimates for scenario from Table 4.6	70
4.6	Boxplots of Interaction Effect standard error / standard deviation for simulation 1	71
4.7	Boxplots of Interaction Effect estimates for scenario from Table 4.8	74
4.8	Boxplots of Interaction Effect standard error / standard deviation for all ICS data	76
5.1	Extension 1: Boxplots of Interaction Effect estimates for Scenario 36 in Table 5.3	84
5.2	Extension 1: Boxplots of Interaction Effect standard error / standard deviation for all ICS data	86
5.3	Extension 2: Boxplots of Interaction Effect estimates for Scenario 1 in Table 5.10	89
5.4	Extension 2: Boxplots of Interaction Effect standard error / standard deviation for Scenario 1 in Table 5.10	90
5.5	Extension 2: Boxplots of Interaction Effect estimates for Scenario 3 in Table 5.13 using smaller SD of cluster size	95

5.6	Extension 2: Boxplots of Interaction Effect estimates for Scenario 3 in Table 5.18 using larger SD of cluster size	96
5.7	Extension 2: Boxplots of Interaction Effect standard error / standard deviation for all ICS data with smaller SD of cluster size	98
5.8	Extension 2: Boxplots of Interaction Effect standard error / standard deviation for all ICS data with larger SD of cluster size	99
5.9	Extension 3: Boxplots of Interaction Effect estimates for Scenario 1 in Table 5.28 for n=60	110
5.10	Extension 3: Boxplots of Interaction Effect standard error / standard deviation for Scenario 1 in Table 5.28 for n=60	111
5.11	Extension 3: Boxplots of Interaction Effect standard error / standard deviation for Scenario 1 in Table 5.30 for n=600	112
5.12	Extension 3: Boxplots of Interaction Effect estimates for Scenario 3 in Table 5.34 for n=60	117
5.13	Extension 3: Boxplots of Interaction Effect estimates for Scenario 3 in Table 5.39 for n=600	118
5.14	Extension 3: Boxplots of Interaction Effect standard error / standard deviation for all ICS data for n=60	121
5.15	Extension 3: Boxplots of Interaction Effect standard error / standard deviation for all ICS data for n=600	122

Abbreviations

ARGEE	Autoregressive Generalised Estimating Equation
CWGEE	Cluster Weighted Generalised Estimating Equation
EGEE	Exchangeable Generalised Estimating Equation
GEE	Generalised Estimating Equation
ICS	Informative Cluster Size
IEE	Independence Estimating Equation
MM	Mixed Model
NICS	Non Informative Cluster Size
POPPET	Providing Optimal Protein for Prems via Enteral Tubes
SD	Standard Deviation of the Estimates
SE	Model Based Standard Error
WCR	Within Cluster Resampling

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: DATE:

Acknowledgements

I would like to begin by thanking my two supervisors, Associate Professor Gary Glonek and Dr. Lisa Yellend, for their help and guidance throughout my Masters.

I am grateful to the staff at the Women's and Children's Health Research Institute who were involved in conducting the POPPET trial and for sharing the data.

Of course I couldn't do this without the help of my family: my dad, Simon, for always being interesting in what I was doing and my mother, Jacqui, for always encouraging me.

Finally, I would like to thanks my partner, Brock, for his ongoing support and always being there for me.

Abstract

Background: Most commonly used statistical methods assume that the data consist of independent observations. Clustered data occur in many settings, such as longitudinal studies, where outcomes are repeatedly measured over time on each subject. Observations from the same subject are dependent and hence form a cluster. Two commonly used methods of analysis for clustered data are mixed models and generalised estimating equations (GEEs).

Additional complexity arises when analysing clustered data where the cluster size is informative; that is, where the cluster size is related to the outcome. Most methods of analysis for clustered data, including mixed models and GEEs, generally assume non informative cluster size and hence may not be suitable when the cluster size is informative.

Aim: The aim of this thesis is to compare methods for analysing longitudinal data when the cluster size (length of follow up) is informative.

Methods: Both real and simulated data were used to compare methods for analysing clustered data with informative cluster size. A range of methods were considered including: GEEs with independent, autoregressive or exchangeable working correlation structures; cluster weighted GEEs; and mixed models. The real data come from a perinatal trial (the POPPET trial), which investigated the effect of high versus standard protein content human milk fortifier on the growth of 60 preterm infants. This dataset was used to investigate different methods of analysis for estimating the effect of treatment on infant growth when informative cluster size was suspected.

As real data cannot be used to show which methods of analysis are performing best in general, a simulation study was conducted to compare methods when the true parameter values were known. The data were simulated based on the POPPET trial. Different treatment effects, sample sizes, and correlations between the cluster size and the outcomes were considered.

Results: For the POPPET trial, evidence of informative cluster size was found. Different methods of analysis produced quite different parameter estimates but similar conclusions about the effect of the intervention.

The simulation results showed that when cluster size was non informative, all methods performed very well. When cluster size was informative, mixed models and autoregressive GEEs always performed well. However, the independence, exchangeable and cluster weighted GEEs often produced low coverage probabilities and model based standard errors that differed from the standard deviation of the parameter estimates. These methods generally performed better when the trial size was larger and when there was no correlation between individual growth trajectories and cluster size.

Conclusions: It is recommended that mixed models or autoregressive GEEs be used to analyse longitudinal data with informative cluster size in general, including the POPPET trial data. Independence, exchangeable and cluster weighted GEEs should only be used when the sample size is large and there is no correlation between individual growth trajectories and cluster size.